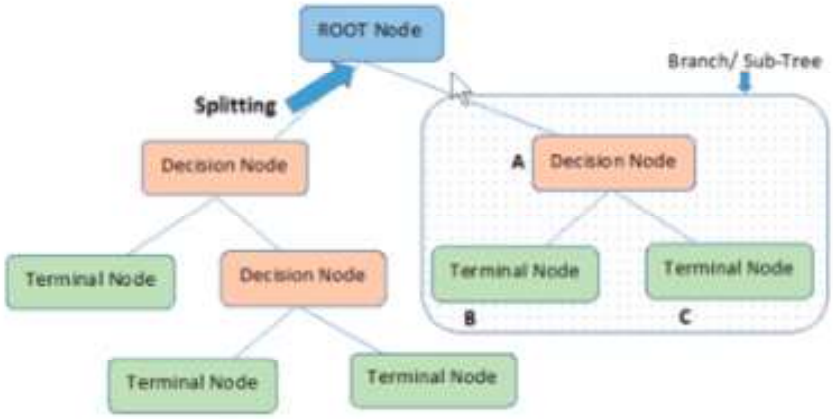
Decision Trees

- Decision Tree model for a supervised learning algorithm which can be used for both classification and regression type of problems.
- We train the model on the training set and validate it on the testing set.
- The parent node gets split into child nodes and pruning is done to avoid overgrowing of sub-trees/branches.

- Mostly used for Classification problems

Logic of Decision Trees

- What if there are more Independent Variables? – Choose the Best Variable using a splitting criterion
- What are the splitting criteria available? - Gini
- How many splits to perform for a data? – Depends on Purity of a Node
- Should we always perform a Binary Split? – Binary & Multiway Split Models are available



Note:- A is parent node of B and C.

Decision Tree – Model Design

- Data should have both 0 (Bad) and 1 (Good) data
- Remove indeterminate values (NAs)
- Look for categorical variables
- Look for meaningful trend. Eg: Height should increase with Age.
- Look for default values like -999. Convert them to missing values. Maybe remove it.
- Look for capping/floor values – Age > 100
- Reasons to create meaningful group – Group all small Northeastern states

For Continuous Independent Variables, different binary cutoff points are chosen and the best Gini Gain cutoff is shortlisted
Such an output is considered as an overfitted Training scenario as it is highly unlikely to maintain the same 100% accuracy on Test data.
However, this example is only to illustrate the concepts of Gini Gain and how it is used to choose the best variable for splitting

Classification Techniques

- **Classification and Regression Tree (CART):**
  - Binary Decision Tree
  - Classification (Categorical output variable)
  - Regression (Continuous output variable)
  - Uses Gini Index
- **CHAID – CHI-squared Automatic Interaction Detector:**
  - Non-Binary Decision Tree
  - Use statistical significance of proportions

Gini Index Calculations

$$Gini\,(D) = 1 - \sum_{i=1}^{m} p_i^{\,2}$$

m : Number of Classes

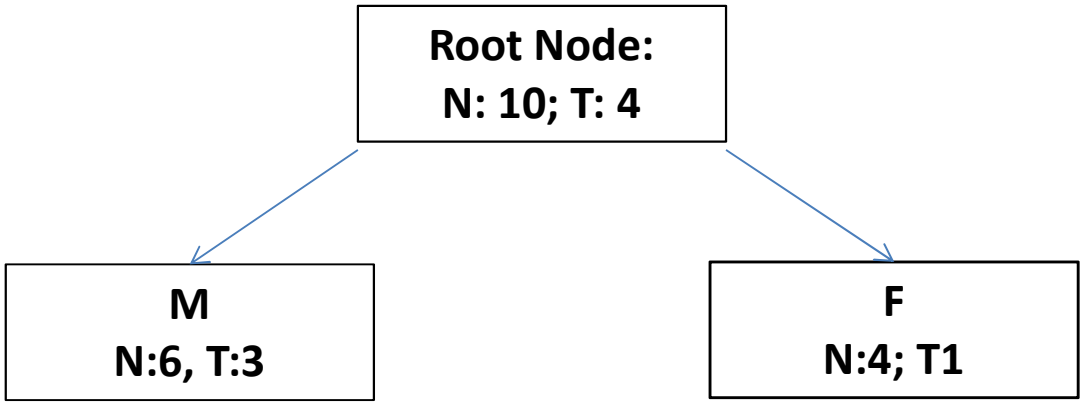
p : Probability that a record in D belongs to class Ci (p → Class Proportion)

Note: The Gini index for the left & the right child nodes has to be calculated. Subtract the child Gini index from the parent Gini index = Gini gain

The independent variable which has the highest Gini gain is said to be the most important variable or the most relevant variable to separate zeros from the ones

GINI Calculation

#	Gender	Decision
1	M	1
2	F	1
3	M	1
4	F	1
5	M	0
6	F	0
7	M	1
8	F	1
9	M	0
10	M	0



Node	GINI Computation Formula	GINI Index
Overall	1 – ((4/10)^2 + (6/10)^2)	0.48
Gender = M	1 – ((3/6)^2 + (3/6)^2)	0.5
Gender = F	1 – ((1/4)^2 + (3/4)^2)	0.375
Gender	(6/10) * 0.5 + (4/10) * 0.375	0.45
GINI Gain	GINI Overall – GINI (Gender)	0.03

GINI Index Values

- ***Gini index varies between values 0 and 1****, where 0 expresses the purity of classification, i.e. All the elements belong to a specified class or only one class exists there.*
- *And 1 indicates the random distribution of elements across various classes.*
- *The value of 0.5 of the Gini Index shows an equal distribution of elements over some classes*

Decision Trees: Advantages & Disadvantages

- Advantages
  - Easy to interpret
  - Automated field selection
  - No data processing required
    - Variable transformation not required
    - Can handle outliers
    - Missing value tolerant
- Disadvantages
  - They are unstable
  - Often inaccurate and poor compared to other models (Solution – Random Forest)
  - Generally not preferred for continuous prediction

Limitations of Decision Trees

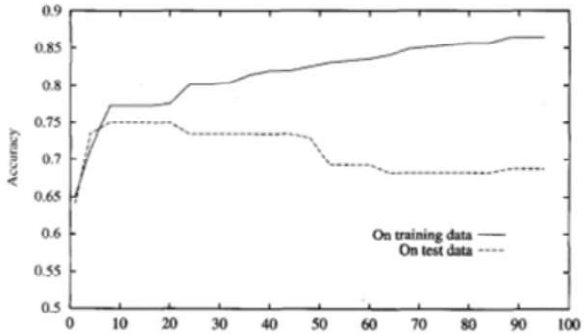
- Vulnerable to over-fitting
  Solution – Pruning

- Greedy Algorithm
  Solution – Cross Validation

Over Fitting & Greedy Algorithm

Over Fitting

- Works extremely well on Training dataset
- Performs poorly on unseen dataset



Greedy Algorithm

Deciding to pay fees on the same day (Greedy Decision, Not Optimal)

Make 31 Paise using any combination of above coins

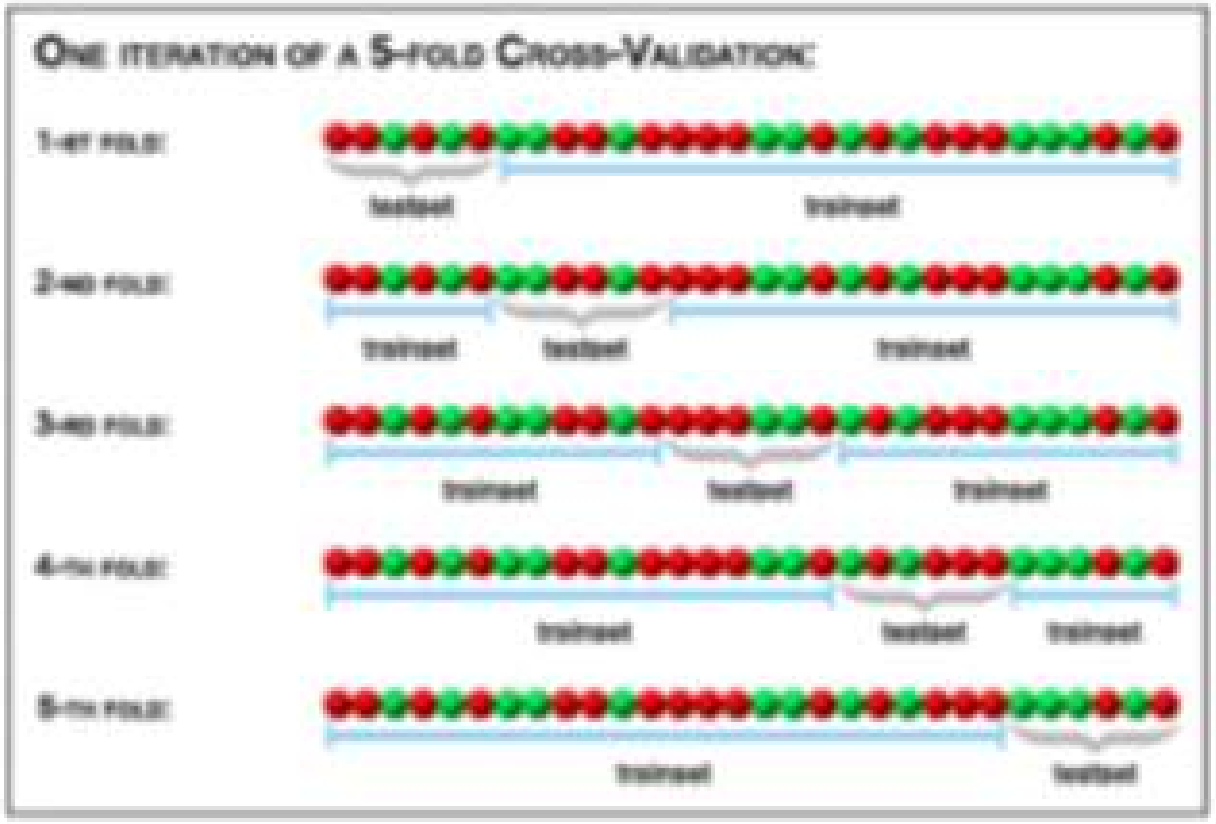
Optimal solution with few coins : 25 + 5 + 1

What if the 5 paise coin is not there?

Optimal solution with few coins : 10 * 3 + 1

Greedy Algorithm solution: 25 + 1 * 6

Cross-Validation



- Helps overcome Greedy Algorithm problem
- How good is the model with unseen data?
- Also helps address 'Over Fitting'

Model Evaluation

Confusion Matrix

a. Confusion Matrix – A 2X2 tabular structure reflecting the performance of the model in four blocks

Confusion Matrix	Predicted Positive	Predicted Negative
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

Type II Error

Type I Error

b. Accuracy – How accurately / cleanly does the model classify the data points. Lesser the false predictions, more the accuracy

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN)$$

c. Sensitivity / Recall – How many of the actual True data points are identified as True data points by the model . Remember, False Negatives are those data points which should have been identified as True.

$$\text{Recall} = TP / TP + FN$$

d. Specificity – How many of the actual Negative data points are identified as negative by the model

$$SPEC = \frac{TN}{TN + FP}$$

e. Precision – Among the points identified as Positive by the model, how many are really Positive

$$\text{Precision} = TP / TP + FP$$

f. F1 Score: (2 * Recall * Precision) / (Recall + Precision)

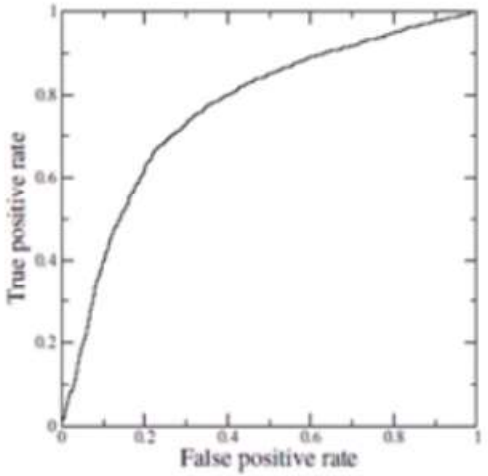
The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0

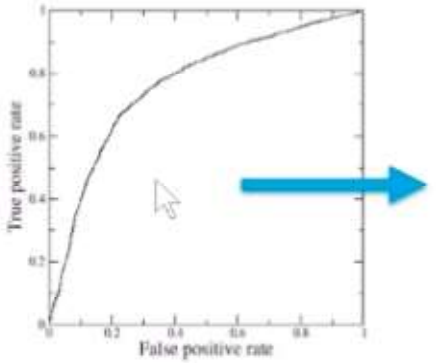
ROC & AUC

Receiver Operating Characteristics (ROC) Curve

A technique for visualizing classifier performance

a. It is a graph between TP rate and FP rates
   I. TP rate = TP / total positive
   II. FP rate = FP / total negative
b. ROC graph is a trade off between benefits (TP) and costs (FP)
c. The point (0,1) represents perfect classified (e.g. D)
   I. TP = 1 and FP = 0
d. Classifiers very close to Y axis and lower (nearer to x axis) are conservative models and strict in classifying positives (low TP rate)
e. Classifiers on top right are liberal in classifying positives hence higher TP rate and FP rate





Area Under the ROC Curve

AUC

Larger the area under the curve, better the model

Handling Imbalanced dataset

	# of customers - PlanA	# of customers - PlanB
# of records	1791	2256
% of records	44%	56%

	Attrition – Yes	Attrition – No
# of records	474	2940
% of records	14%	86%

	Cancer - Yes	Cancer – No
# of records	152	12346
% of records	1.2%	98.8%

	Fraudulent transactions - Yes	Fraudulent transactions - No
# of records	84	12492
% of records	0.67%	99.33%

Techniques to Balance the data:

- **Synthetic Minority Oversampling Technique**, or **SMOTE**
- **Under Sampling**
- **Over Sampling**